AD A107817

LEVEL II

(20)

RESEARCH REPORT

PJH - 6

PROJECTION PURSUIT

Peter J. Huber
August 1981

This work was facilitated in part by National Science Foundation Grant MCS-79-08685 and Office of Naval Research Contract N00014-79-C-0512.

Department of Statistics
Harvard University
Cambridge

PROJECTION PURSUIT

Peter J. Huber
Harvard University

ABSTRACT

Projection pursuit (PP), which goes back to Kruskal ~~(1969)~~ and Friedman and Tukey (1974), is discussed from a conceptual point of view. Originally, it was merely concerned with finding "interesting" projections to aid the visual analysis of high dimensional data. It will be argued that there is a very general concept behind PP, having a much broader scope than anticipated by its originators. In its most important aspect it is concerned with finding least normal projections of the data, and it has ramifications reaching into topics as varied as robust covariance estimation, factor analysis, nonparametric signal detection, computer tomography and Hilbert's 13th problem.

Best
Available
Copy

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

TABLE OF CONTENTS

INTRODUCTION

Projection pursuit techniques were proposed by Kruskal (1969) and implemented by Friedman and Tukey (1974), (who coined the catchy name), to aid the visual analysis of high-dimensional point clouds. The original purpose of PP was to machine-pick "interesting" low-dimensional projections by numerically maximizing a certain projection index. In its original form, this index was the product of a robust measure of scale (trimmed standard deviation) with a measure of lumpiness (a weighted count of the number of close pairs in the projection).

Later on, after a dormant stage of several years, Friedman and his coworkers extended the idea behind PP and added projection pursuit regression (PPR; Friedman, Jacobson and Stuetzle 1980) and projection pursuit classification (PPC; Friedman and Stuetzle 1980).

The most exciting feature of PP is that PP is one of the very few multivariate methods able to bypass the "curse of dimensionality," caused by the fact that high dimensional space is mostly empty. For example, if a large number of points is distributed uniformly in the 10-dimensional ball, then the radius of a ball containing 5% of the points is $0.05^{1/10} \approx 0.74$. As a consequence, methods like kernel smoothers will not be able to pick up small features, unless the sample size is truly gigantic. PP avoids the problem by working in low-dimensional linear projections. The price to be paid is, of course, that PP is poorly suited to deal with highly nonlinear structures (but kernel smoothers then are not a viable alternative either).

In addition, some PP methods are able to ignore redundant (i.e., noisy and information-poor) variables. This can be a distinct advantage over methods based on interpoint distances, like minimal spanning trees, multidimensional scaling and most clustering techniques. But note that the original PP of Friedman and Tukey does not have this property.

This author suggested in an unpublished draft (1979) that the purpose of PP, namely looking for interesting projections, might be formalized as a search for least normal projections, but his approach, based on maximizing estimated Fisher information, foundered because of troubles with inefficient numerical estimation and optimization procedures. Donoho (1980) was quick to pick up the idea; he pointed out that minimum entropy deconvolution (MED), as used in the analysis of time series data in oil exploration, is a special case of PP, and the goal there indeed is finding least normal projections.

The quartimax method in factor analysis, see Harman (1967, ch. 14), is another special case of PP.

Among the more remote ramifications of PP one should mention Computer Assisted Tomography (CAT): both PP and CAT are concerned with the efficient analysis of higher dimensional data through the use of lower dimensional projections, and there may be interesting cross-fertilizations.

Furthermore, there is an amusing connection between PPR and Hilbert's 13th Problem.

The present study attempts to analyze the PP idea by taking it apart into its conceptual components. We hope that this will clarify the goals of PP, will lead to a better understanding of how it is supposed

to work, will suggest new applications, and will help to improve the procedures.

In particular, we shall separate PP into an "abstract" version operating on random variables and probability densities, and its "practical" implementations, operating on finite point clouds.

The "problems" to be mentioned in the following sections are open research problems; some of them may be trivial, some unsolvable.

2. PRINCIPLES OF PP

2.1 Generalities

We begin with a kind of "philosophical" analysis of PP; we shall enunciate a number of theses, each followed by a heuristic motivation and explanation, and often preceded by some definitions.

<u>THESIS 1</u>. PP helps to elucidate an underlying structure (of which the observed point-cloud is a sample).

Thus, we separate PP into an "abstract" version operating on p-dimensional distributions (mostly: probability densities in $\mathbb{R}^p$), and a "practical" version that is applied to samples (i.e., empirical distributions of "point-clouds"). The two versions might be identical, but often, the abstract version will work on smooth distributions only, so, in order to translate it into a practical one, we must insert a suitable smoother at the appropriate place. Initially, we shall only be concerned with the abstract version, and we shall postpone questions of smoothing and of sampling properties.

A <u>linear projection</u> from $\mathbb{R}^p$ onto $\mathbb{R}^k$ is any linear map A, or $k \times p$ matrix, of rank k:

$$z = Ax, \qquad x \in \mathbb{R}^p, \qquad z \in \mathbb{R}^k.$$

We speak of an <u>orthogonal projection</u>, if the row vectors of A are orthogonal to each other and have length 1.

If x is a p-dimensional random variable with distribution F, then $z = Ax$ is a k-dimensional random variable with distribution F_A. If $k = 1$, A reduces to a row vector a^T, and we then use lower case letters: F_a, etc.

In passing, we note that any p-dimensional distribution F is identifiable by its one-dimensional projections F_a. This follows trivially from the fact that F is uniquely characterized by its characteristic function ψ, and that the characteristic function ψ_a of the one-dimensional projection F_a in direction $\underset{\sim}{a}$ equals the section of ψ along the same direction:

$$\psi_a(t) = E(e^{ita^T x}) = \psi(ta).$$

By definition, PP searches for a projection A maximizing (or minimizing) a certain <u>projection index</u> or <u>objective function</u> $Q(F_A)$. We are specifically interested not only in absolute, but also in local extrema. While Q is a functional on the space of distributions on $\mathbb{R}^k$, we find it more convenient to use random variable terminology and, by abuse of notation, to write $Q(Ax)$ instead of $Q(F_A)$. Primarily, we shall be concerned with one-dimensional projections, and for obvious (representational) reasons we shall rarely want to go beyond three-dimensional projections.

Over what domain should A range, and what constitutes an "interesting" projection, that is, how should Q be chosen?

Clearly, the <u>coordinate directions</u> deserve special scrutiny, so almost every data analysis will start with a visual inspection of the

orthogonal projections on the spaces spanned by one or two, perhaps three, coordinate axes. If there are many variables, one cannot of course look at all pairs or triples, and one will inspect only a selection consisting of those with the largest projection index (i.e., PP can be used as a method for variable selection).

2.2 Principal Components

Next, there are the directions of the leading *principal components* (the eigenvectors belonging to the largest eigenvalues of the covariance matrix, or of a robustified version thereof). Often, they show interesting structure. Why? There seem to be at least two somewhat different reasons.

First, if a population is an aggregate of several clusters, then these clusters can become individually visible only if the separations between them are larger than the internal scatter of the clusters. Thus, if there are only few clusters, the leading principal axes will tend to pick projections with good separations. Of course, principal components can go astray if there too many, isotropically distributed clusters (compare the Friedman-Tukey (1974) example where the clusters are at the corners of a regular simplex), or if there are meaningless variables with a high noise level.

The second reason is more germane to the principal component analyses performed on correlation matrices: assume that we have an intrinsic structure describable by a few (unobservable) variables, and that we observe many, possibly differently scaled (linear) functions

of these variables, with independent random noise added. Then principal component analysis tends to act as a variation reducing technique (not unlike the sample mean), relegating most of the random noise to the trailing components, and collecting the systematic structure into the leading ones.

We note that both the mean vector $\mu = \mathrm{ave}(x)$ and the principal components, i.e. the eigenvalue/eigenvector representation of the covariance matrix $\Sigma = \mathrm{ave}\{(x-\mu)(x-\mu)^T\}$ can be captured by PP methods, see the following examples.

Example 2.1. Define the objective function $Q(a^T x) = \mathrm{ave}(a^T x)$ with $\|a\| = 1$. This is maximized by $a_0 = \mu/\|\mu\|$, and the value at the maximum is $Q(a_0^T x) = \|\mu\|$.

Example 2.2. Let $Q(a^T x) = \mathrm{ave}\{[a^T(x-\mu)]^2\}$ with $\|a\| = 1$. The maximum value of this objective function is the largest eigenvalue of Σ, and it is reached at any eigenvector belonging to this eigenvalue. The other eigenvalues and eigenvectors can be found successively by restricting a to the orthogonal complement of the space spanned by the previously found eigenvectors.

Thus, principal components analysis is a special case of PP. The PP approach suggests interesting variations. For instance, if we replace the object function of Example 2.2 by any robust measure of scale, we obtain a robust version of principal component analysis (see Chen and Li (1981), and the following section 2.3).

2.3 Objective Functions Classified According to Invariance

We single out a few classes of objective functions according to their invariance properties. For simplicity, we consider only one-dimensional orthogonal projections but the ideas generalize.

Let z be a real random variable, while a, t denote (nonrandom) real numbers. We distinguish three classes of objective functions Q:

<u>CLASS I.</u> Location-scale equivariance:

$$Q_I(az+t) = aQ_I(z) + t.$$

<u>CLASS II.</u> Location invariance, scale equivariance:

$$Q_{II}(az+t) = |a|Q_{II}(z).$$

<u>CLASS III.</u> Affine invariance:

$$Q_{III}(az+t) = Q_{III}(z), \quad a \neq 0.$$

We note that Q_I is a location functional, and that PP with a Class I objective function yields a kind of p-dimensional location estimate (we say "kind of" because it is not fully location equivariant in general; also, it need not be uniquely defined). In somewhat more detail this works as follows. Assume that $a_o \in \mathbb{R}^p$, with $\|a_o\| = 1$ maximizes $Q_I(a^T x)$ and put $T(x) = a_o\, Q_I(a_o^T x)$.

<u>PROPOSITION 2.3.1.</u> The functional T is uniquely defined and location equivariant for the translation family generated by $\mathcal{L}(x)$:

$$T(x+t) = T(x) + t \text{ for all } t \in \mathbb{R}^p,$$

iff there is a $\mu \in \mathbb{R}^p$ such that

$$Q_I(a^T(x-\mu)) = 0 \text{ for all } a \in \mathbb{R}^p, \qquad (*)$$

and then $T(x) = \mu$. In particular, this condition holds if the distribution of x is centrosymmetric about μ (i.e., if $x-\mu$ and $-(x-\mu)$ have the same distribution).

<u>Proof.</u> If the distribution of x is symmetric about μ, then the distribution of $z = a^T(x-\mu)$ is symmetric about 0, and

$$Q_I(-z) = -Q_I(z) = Q_I(z) = 0,$$

which establishes the last statement.

The condition (*) is sufficient: if it holds, then

$Q_I(a^T(x+t)) = Q_I(a^T(x-\mu) + a^T(\mu+t)) = Q_I(a^T(x-\mu)) + a^T(\mu+t) = a^T(\mu+t)$

which is maximized for $a = a_o = (\mu+t)/\|\mu+t\|$, with $Q_I(a_o^T(x+t)) = \|\mu+t\|$; hence $T(x+t) = \mu+t$.

Conversely, if T is translation equivariant, put $\mu = T(x)$. Take an arbitrary fixed value $t \in \mathbb{R}^p$, and let $a_o = T(x+t)/\|T(x+t)\| = (\mu+t)/\|\mu+t\|$. Then

$$\begin{aligned}\sup Q_I(a^T(x+t)) &= \|T(x+t)\| = \|T(x) + t\| = \|\mu+t\| \\ &= \sup Q_I(a^T(x-\mu) + a^T(\mu+t)) \\ &= Q_I(a_o^T(x-\mu)) + a_o^T(\mu+t) \\ &= Q_I(a_o^T(x-\mu)) + \|\mu+t\|,\end{aligned}$$

hence $\quad Q_I(a_o^T(x-\mu)) = 0.$

Since t was arbitrary, it follows that $Q_I(a^T(x-\mu)) = 0$ for all $a \in \mathbb{R}^p$.

Clearly, if the condition of Proposition 2.3.1 holds, the estimate T is uniquely defined as $T(x+t) = \mu+t$. ||

PROBLEM 1. Find the class of Q_I for which PP gives a unique p-dimensional estimate for arbitrary distributions $(p > 1)$, and find the subclass for which this estimate is location equivariant. Are there any nontrivial Q_I of this kind (i.e., different from Example 2.1.1)?

PROBLEM 2. Investigate the properties, in particular the robustness properties of the (not necessarily location equivariant nor unique) location estimates defined in the above way.

Any Class II functional yields a generalized principal component decomposition and a pseudo-covariance matrix in the manner sketched in Section 2.2.

PROBLEM 3. Investigate the properties, in particular the robustness properties of covariance estimates defined through PP and a Class II functional. For first steps in this direction, see Chen and Li (1981).

The attractiveness of multivariate estimates derived in this way lies in the observation that they inherit certain properties, for instance their breakdown point, from the underlying one-dimensional location or scale functional Q_I or Q_{II} respectively. For example, a PP covariance estimate based on the median absolute deviation achieves a breakdown point of 0.5, and is orthogonally equivariant (i.e. commutes with orthogonal transformations).

For Class III functionals, the PP estimate is affinely equivariant (i.e., the solution commutes with affine transformations). So, in particular, the distinction between linear (more precisely, affine) and orthogonal projections disappear. Technically, this means that we may use unconstrained optimization (i.e., without imposing the constraint $\|a\| = 1$).

PP with non-affinely-invariant objective functions clearly gives rise to interesting problems and has interesting applications; the most important of them seem to be generalizations of principal components analysis, like the examples above and the original Friedman-Tukey PP.

But in the affinely invariant case (with Class III functionals) some novel and most intriguing features emerge. We formulate this in the following thesis:

THESIS 11. The most important new feature of PP methods is that they can pick features of multivariate distributions not obtainable through mere knowledge of the mean vector and covariance matrix, and if so desired, they can clearly separate these features from the information contained in mean and covariance matrix.

Indeed, PP based on an affinely invariant (i.e., Class III) functional ignores the information contained in the mean vector and the covariance matrix.

Instead of a Class III functional, we may also take the following--perhaps somewhat more intuitive--approach. We first standardize the data by subtracting an affinely equivariant location estimate μ and by removing the covariance structure, e.g., by multiplying the data by a lower triangular matrix L:

$$y = L\,(x-\mu),$$

where $\Sigma = L^{-1}L^{-T}$ is the Cholesky decomposition of the covariance matrix Σ of the x. (Perhaps we may prefer to use a robust, affinely equivariant version of Σ.)

Then, we use any orthogonally invariant objective function on the transformed data y, with orthogonal PP (in dimension 1, orthogonal invariance simply means $Q(-a) = Q(a)$).

It is intuitively obvious that this approach yields a version of PP whose results are equivariant under affine transformation of the original data x. The formal proof of this statement is slightly tricky because of the possible non-uniqueness of the solutions; we define (affine) equivariance by requiring that the transform of any solution based on the original data is a solution of the problem based on the transformed data.

Thus, if x is affinely transformed into $\tilde{x} = Sx + v$, then $\tilde{\mu} = S\mu + v$ and $\tilde{\Sigma} = S\Sigma S^T$ are solutions of the transformed problem (not necessarily the only ones), and $\tilde{L} = ULS^{-1}$ for some orthogonal matrix U (chosen such as to make $\tilde{L}$ lower triangular). Thus

$$\tilde{y} = \tilde{L}(\tilde{x} - \tilde{\mu}) = ULS^{-1}(Sx + v - S\mu - v) = Uy,$$

and the extrema of $Q(a^T y)$ and $Q(\tilde{a}^T \tilde{y})$ respectively are reached at directions related by $\tilde{a} = Ua$ and have the same values.

A strong motivation for Thesis II, that is for looking beyond mean and covariance, and for separating out these aspects, is contained in the following considerations.

First, we know that a p-variate normal distribution is completely specified by its mean vector and covariance matrix, and furthermore, sample

mean and covariance then are sufficient statistics. In other words, in a guaranteed multivariate normal situation, there is no need to go beyond mean and covariance. In particular, we need not look at projections: all projections are normal, with means and covariance matrices computable form the p-dimensional mean and covariance matrix.

Conversely, if all one-dimensional projections are normal, then we have multivariate normality (this is one of the well known characteristics of multivariate normality).

Second, if the dimensionality p is high, and if the coordinates of x are approximately independent (at least in a suitable coordinate system), then it follows from the central limit theorem that most projections are nearly normal (of course, subject to certain regularity conditions, e.g., bounds on the standardized third absolute moment).

This leads us into formulating the following extension of Thesis II:

<u>THESIS III</u>. The projections most interesting for visual inspection are the least normal ones, and therefore, we should devise PP methods specifically for picking least normal projections.

3. LEAST NORMAL PROJECTIONS

In this section, we exclusively consider one-dimensional projections, and we assume that all our Q are of Class III, i.e., affinely invariant.

Now, taking off from Thesis III, consider the simplest case where the coordinates $x_1, \ldots, x_p$ of x are independent and have the same non-normal distribution with finite, non-zero variance. Then it is intuitively plausible that Q should be such that $Q(x_i) \geq Q(\sum a_i x_i)$, since $\sum a_i x_i$ is "more normal" than any single x_i. In other words, in this case PP should pick the pure coordinate directions.

Following [illegible] (1980), we introduce a partial order among non-degenerate distributions with finite variance on $\mathbb{R}$, or more precisely, among equivalent classes of non-degenerate random variables.

DEFINITION 3.1. Two non-degenerate random variables X and Y are equivalent, $X \doteq Y$, if $\mathcal{L}(Y) = \mathcal{L}(aX + b)$ for some real numbers a, b.

DEFINITION 3.2. We write $X \preceq Y$ if $\mathcal{L}(Y) = \mathcal{L}(\sum_1^n a_i X_i + b)$ for some n, b and some non-zero a_i, where the X_i are independent copies of X.

It is clear that $\preceq$ is transitive. It follows from Theorem 5.6.1 of Kagan, Linnik and Rao (1973) that

$$X \preceq Y \text{ and } Y \preceq X \implies X \doteq Y. \qquad (*)$$

THESIS IV. If we are interested in finding least normal projections, we should use objective functions that preserve order:

$$X \preceq Y \implies Q(X) \geq Q(Y).$$

Note that this requirement implies affine invariance: $Q(aX+b) = Q(X)$, hence Q is of Class III. Moreover, if Q is weakly lower semi continuous, it follows that Q reaches its minimum at the normal distribution.

Proof. Let X_i be i.i.d. with zero expectation and finite variance, and put $Y_n = n^{-1/2} \sum_1^n X_i$. Then $Y_1 \preceq Y_2 \preceq \ldots \preceq Y_{2^j} \preceq \ldots$, and $\mathcal{L}(Y_{2^j})$ converges weakly to a normal distribution $\mathcal{L}(Y)$ by the central limit theorem. Because of lower semi-continuity of Q it now follows that $\lim Q(Y_{2^j}) \geq Q(Y)$, hence $Q(X_1) \geq Q(Y)$. ‖

We note that the order $\preceq$ has too many incomparable classes to be entirely satisfactory. For example, the normal distribution is not comparable to any other class: if Y is normal, then $Y \preceq X$ clearly implies that X is normal, and conversely, $X \preceq Y$ implies normality of X by a well-known characterization theorem for the normal law. This remark suggests the following research problem.

PROBLEM 4. Investigate the properties of the relation $\preceq$ defined by putting $X \preceq Y$ if the inequality $Q(X) \geq Q(Y)$ holds for all lower semi-continuous Q satisfying Thesis IV.

Upon second thought, it appears that we would prefer a slightly stronger property than the one enunciated in Thesis IV, namely

THESIS IVa. Q should be affinely invariant, and if X, Y are independent then

$$Q(X+Y) \le \max(Q(X), Q(Y)).$$

It follows at once by induction that IVa implies IV.

EXAMPLE 1.1. Let $c_m(X) = \left[(-i)^m \frac{d^m}{dt^m} \log E(e^{itX})\right]_{t=0}$ be the mth cumulant of the real random variable X, then $c_m(X+Y) = c_m(X) + c_m(Y)$ if X and Y are independent. Furthermore, let $\gamma_m(X) = c_m(X)\, c_2(X)^{-m/2}$ be the standardized cumulant, $m > 2$. Note that

$$\gamma_m(X) = c_m\left(\frac{X}{\sigma(X)}\right).$$

Then

$$|\gamma_m(X+Y)| = \left|c_m\left(\frac{X+Y}{\sigma(X+Y)}\right)\right|$$

$$= \left|c_m\left(\frac{X}{\sigma(X+Y)}\right) + c_m\left(\frac{Y}{\sigma(X+Y)}\right)\right|$$

$$= \left|\left[\frac{\sigma^2(X)}{\sigma^2(X+Y)}\right]^{m/2} \gamma_m(X) + \left[\frac{\sigma^2(Y)}{\sigma^2(X+Y)}\right]^{m/2} \gamma_m(Y)\right|$$

$$\le \left\{\left[\frac{\sigma^2(X)}{\sigma^2(X+Y)}\right]^{m/2} + \left[\frac{\sigma^2(Y)}{\sigma^2(X+Y)}\right]^{m/2}\right\} \max\{|\gamma_m(X)|, |\gamma_m(Y)|\}$$

$$\le \max\{|\gamma_m(X)|, |\gamma_m(Y)|\}.$$

It follows that $Q(X) = |\gamma_m(X)|$, $m > 2$, satisfies the requirement of Thesis IVa. We recall Ferguson's result (1961) that skewness (γ_3) and kurtosis (γ_4) are, in a certain sense, the best outlier tests for an underlying normal model, especially if the number of outliers is not specified in advance. Thus, if we want to develop specific PP methods for finding multivariate outliers, objective functions (γ_3) and (γ_4) would be the leading contenders.

NOTE. The quartimax method of factor analysis (cf. Harman 1967, Ch. 14) amounts to PP based on kurtosis. Thus, the above arguments simultaneously give a theoretical foundation for the quartimax method, and point out a major weakness (outlier proneness).

EXAMPLE 1.2. Let $E_{sh}(X) = E_{sh}(f) = -\int \log(f)\, f\, dx$ be Shannon entropy, and put $Q(X) = -E_{sh}(X) + \log(\sigma(X))$ where $\sigma^2(X)$ is the variance of X. Then Q is affinely invariant and satisfies Thesis IVa. Moreover, the normal distribution is uniquely characterized by the property that it minimizes Q.

Proof. Invariance under translation is trivial to show. If we put $f_\sigma(x) = \frac{1}{\sigma} f(x/\sigma)$, then

$$E_{sh}(f_\sigma) = \int \{\log \sigma - \log f(x/\sigma)\} \frac{1}{\sigma} f(x/\sigma)\, dx$$

$$= \log \sigma + E_{sh}(f),$$

hence Q is also invariant under scale changes.

If f has variance 1 and minimizes Q or $-E_{sh}$, then it satisfies the following three variational conditions:

$$\delta \int \log(f)\, f\, dx = \int (1 + \log f)\, \delta f\, dx = 0,$$

$$\delta \int x^2 f\, dx = \int x^2\, \delta f\, dx = 0,$$

$$\delta \int f\, dx = \int \delta f\, dx = 0.$$

If we combine them with the aid of Lagrange multipliers, we find that

$$\log f(x) = -ax^2 - b,$$

so f must be normal.

The proof that Thesis IVa is satisfied proceeds as follows. We first note that

$$E_{sh}(sX) = E_{sh}(X) + \log s,$$

and hence

$$Q(X) = -E_{sh}\left(\frac{X}{\sigma(X)}\right).$$

If X and Y are non-degenerate independent random variables, then

$$e^{2E_{sh}(X+Y)} \geq e^{2E_{sh}(X)} + e^{2E_{sh}(Y)},$$

with equality only if X and Y are normally distributed, see Blachman (1965).

In particular, we have

$$\exp\left\{2E_{sh}\left(\frac{X+Y}{\sigma(X+Y)}\right)\right\} \geq \exp\left\{2E_{sh}\left(\frac{X}{\sigma(X+Y)}\right)\right\} + \exp\left\{2E_{sh}\left(\frac{Y}{\sigma(X+Y)}\right)\right\}$$

$$= \exp\left\{2E_{sh}\left(\frac{X}{\sigma(X)}\right)\right\} \frac{\sigma^2(X)}{\sigma^2(X+Y)} + \exp\left\{2E_{sh}\left(\frac{Y}{\sigma(Y)}\right)\right\} \frac{\sigma^2(Y)}{\sigma^2(X+Y)}$$

$$\geq \frac{\sigma^2(X)+\sigma^2(Y)}{\sigma^2(X+Y)} \min\left[\exp\left\{2E_{sh}\left(\frac{X}{\sigma(X)}\right)\right\}, \exp\left\{2E_{sh}\left(\frac{Y}{\sigma(Y)}\right)\right\}\right]$$

Since the factor in front of the min is 1, it follows that Q satisfies Thesis IVa, and that the inequality is strict, unless X and Y are both normal. (Proof suggested by D. Donoho.)

<u>EXAMPLE 3.3.</u> Let

$$I(X) = I(f) = \int (f'/f)^2 f \, dx$$

be Fisher information, and put

$$Q(X) = \sigma^2(X)\, I(X)$$

Then Q is affinely invariant, and satisfies Thesis IV:

$$Q\left\{ \sum a_i X_i \right\} \le Q(X).$$

Moreover, the normal distribution is uniquely characterized by minimizing Q.

<u>Proof.</u> Let $f_\sigma(x) = \frac{1}{\sigma} f(\frac{x}{\sigma})$, then

$$I(f_\sigma) = \int \frac{1}{\sigma^2} \left(\frac{f'(x/\sigma)}{f(x/\sigma)} \right)^2 \frac{1}{\sigma} f(\frac{x}{\sigma})\, dx = \frac{1}{\sigma^2} I(f).$$

It follows that Q is invariant under scale transformations (and clearly also under translations). One easily verifies that the normal distribution satisfies the variational conditions for a minimum of Q. Uniqueness follows easily from convexity of Fisher information; for details, see Huber (1981a), p. 80ff.

Thesis IV can be verified neatly in the special case where the a_i are equal. We first note that for any k, the best location estimate based on $Y_1, \ldots, Y_k$, with $Y_i = (X_{(i-1)p+1} + \ldots + X_{ip})/p$ cannot be any better than the best estimate based on $X_1, \ldots, X_{kp}$. Asymptotically for $k \to \infty$, the best estimates will be normal with variances equal to the respective Cramér-Rao bounds and it follows that there must a certain inequality, equivalent to Thesis IV, between these bounds.

The proof that Thesis IVa is satisfied proceeds as follows. We note that

$$I(aX) = I(X)/a^2,$$

and that

$$Q(X) = I\left(\frac{X}{\sigma(X)}\right).$$

If X and Y are independent, then

$$\frac{1}{I(X+Y)} \ge \frac{1}{I(X)} + \frac{1}{I(Y)},$$

with equality only if X and Y are both normal, see Blachman (1965). In particular, we have

$$\frac{1}{I\left[\frac{X+Y}{\sigma(X+Y)}\right]} \ge \frac{1}{I\left[\frac{X}{\sigma(X+Y)}\right]} + \frac{1}{I\left[\frac{Y}{\sigma(X+Y)}\right]}$$

$$= \frac{\sigma^2(X)}{\sigma^2(X+Y)} \frac{1}{I\left[\frac{X}{\sigma(X)}\right]} + \frac{\sigma^2(Y)}{\sigma^2(X+Y)} \frac{1}{I\left[\frac{Y}{\sigma(Y)}\right]}$$

$$\ge \frac{\sigma^2(X) + \sigma^2(Y)}{\sigma^2(X+Y)} \min\left\{ \frac{1}{I\left[\frac{X}{\sigma(X)}\right]}, \frac{1}{I\left[\frac{Y}{\sigma(Y)}\right]} \right\}$$

Since the factor in front of the min is 1, it follows that Q satisfies Thesis IVa, and that the inequality is strict, unless X and Y are both normal.

EXAMPLE 3.4. Assume that Q satisfies Thesis IVa. For any random variable X with finite variance put $X^* = X + c\,\sigma(X)\,Z$, where Z is a standard normal random variable independent of X, $\sigma(X)$ is the standard deviation of X, and c is a positive real number. Define

$$Q^*(X) = Q(X^*).$$

Then Q^* satisfies Thesis IVa; this follows rather trivially from the remark that $(X+Y)^* = X^* + Y^*$. The importance of this example lies in the fact that Q^* is defined for discrete, in particular, empirical distributions: it consists in applying Q to the distribution of X smoothed by a normal kernel.

In passing, we note that the nonnormality of the least normal marginal distribution can be used as a measure of the nonnormality of the joint distribution. This leads to attractive proposals for tests of joint normality.

PROBLEM 5. Investigate the properties of such tests (they will depend on the objective function Q, possible on the smoothing methods used, and finally on a particular test for one-dimensional normality).

At the end of this section, a word of caution should be added: long tails, i.e., outliers, constitute one of the most easily detectable types of deviation from normality, yet they usually are uninteresting in the sense that they are not generated by the underlying structure (the object of investigation) but by inadequacies of the measuring and recording process. Therefore, we should either spot outliers and trim the tails (possibly aided by PP methods) before the PP structure search is done, or we should design the objective functions such that their sensitivity to outliers is lowered. The latter approach ordinarily will lead to objective functions violating Thesis IV.

4. QUESTIONS OF k-DIMENSIONAL PROJECTIONS

In the preceding section we were concerned with one-dimensional projections. The same approach, that is, maximizing some functional Q of distributions on $\mathbb{R}^k$, applies in higher dimensions, but it has drawbacks:

(1) computations get harder (minimization over approximately kp instead of p variables);

(2) it yields only a k-dimensional subspace, but for interpretational reasons one would like to get an ordered set of k directions.

Therefore, stepwise approaches looks attractive: fix the first k-1 directions found and optimize among projections onto the k-space spanned by the fixed k-1 plus one additional variable direction. But also this does not give a sequence of directions, only a nested sequence of subspaces.

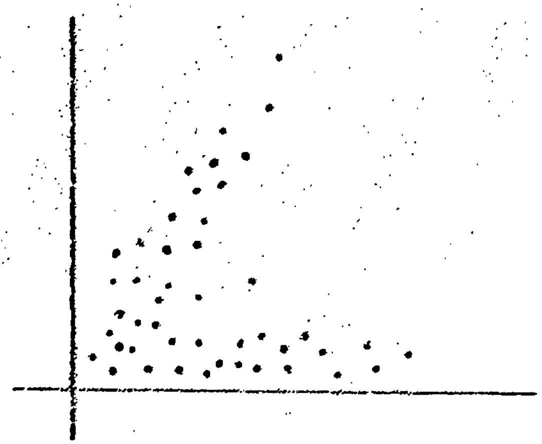

FIGURE 4.1

Note that orthogonal directions do not suffice, the interesting directions may be oblique to each other. When doing manual projection pursuit with actual data some months ago, we met a *few* striking pictures of the type sketched in Figure 4.1, where the points concentrate along two oblique directions. (It is unfortunate that we did not have an operating hardcopy device at that time.)

Of course, the approach can be reversed: find first a k-dimensional projection, then reduce dimension one by one.

<u>PROBLEM 6</u>. Investigate the advantages and disadvantages of stepwise procedures relative to direct k-dimensional projections.

5. WHAT NEXT?

After one has found one or more "interesting" projections, what does one do next? Usually, the next action is one of the following list (parts (1) and (3) roughly correspond to PPC and PPR, see Section 1, and below):

(1) Identify clusters, isolate them and investigate them separately.

(2) Identify clusters and locate them (i.e., replace them by their center and classify points according to their membership to a cluster).

(3) Find a parsimonious description (separate structure from random noise in a nonparametric fashion).

Clearly, there is a floating boundary between the entries in this list, and the details need further investigation.

We note that often a cluster can be characterized by the location of its center and the scatter matrix of the points forming the cluster.

Assume for the moment that we would like to optimize a PP procedure for finding clusters. Then, even in the relatively simple case of overlapping elliptical clusters with discrete centers, it is far from clear how we should optimize the choice of objective functions Q. In view of Theorem III, the problem of detecting such clusters is equivalent to a test of normality whose power is optimized for a particular class of nonparametric alternatives. Clearly, PP here infuses some new life and problems

into the old area of nonparametric normality tests. On the other hand, determining the shape of the clusters is a problem of robust estimation of scatter matrices.

If the clusters are no longer elliptical, a description in terms of scatter matrices may become inappropriate. For nonconvex clusters (e.g., for curved "sausages") low dimensional projections should still be able to reveal the presence of structure, but they may be of little help in unravelling it, mainly because each projection may show confusing overlapping effects.

In such cases, a separation of structure from noise ("sharpening") may reduce overlapping effects and thus help with the interpretation. It recently has emerged that PP methods are able to yield one of the most general and theoretically cleanest approaches to sharpening by deconvoluting the underlying distribution (see Section 7). But first we must discuss some representational problems.

6. REPRESENTATIONAL PROBLEMS

Assume that we want to approximate (or, in the limit, represent) a function f of many variables $x_1,\ldots,x_p$. In concrete terms, f may be a density, or a response surface.

In one dimension, the most common approach is to expand f into a series

$$f(x) = \sum a_j \phi_j(x),$$

where the basis functions ϕ_j might be (not necessarily orthogonal) polynomials, trigonometric functions, B-splines, etc.

In p dimensions, the obvious generalization is to use the Kronecker product of the one-dimensional bases as a basis in p-space, formed by the functions

$$\phi_{j_1 j_2 \ldots j_p} = \phi_{j_1}(x_1)\,\phi_{j_2}(x_2)\,\ldots\,\phi_{j_p}(x_p).$$

However, this runs head-on into the curse of dimensionality. In a statistical context we might need between 5^p and 10^p observations to determine a meaningful number of coefficients in such a series expansion of f.

A possible way out is recursive partitioning (cf. J. Friedman 1979): split the domain of f into parts, always splitting into two that part which contains most of the action, until the function can satisfactorily be approximated on each part by, say a linear function.

It is intuitively plausible that this should work better for intrinsically wild functions, pieced together from unrelated functions defined on some tessellation of the domain, than for intrinsically smooth functions. In statistics one would seem to be particularly interested in approaches generalizing and extending the traditional linear models, and therefore, while recursive partitioning may be very well be adapted for classification purposes, it does not look so attractive in the regression context.

But approaches based on PP do: represent f by a (generally infinite) sum

$$f(x) = \sum f_j\,(a_j^T x) \qquad (*)$$

of functions f_j of a single real parameter each, applied to suitable one-dimensional projection, or approximate f by a finite such sum. Such an approach, called projection pursuit regression (PPR) was first proposed by Friedman, Jacobson and Stuetzle (1980).

We begin with a few simple theoretical considerations. First, in which sense should the series (*) converge to f? If the domain of x is unbounded, then the summands are not Lebesgue integrable, so, say, L_2-convergence makes only sense with respect to a bounded (i.e. probability) measure P in $\mathbb{R}^p$.

To fix the idea, we may take P to be the uniform measure on the unit cube in $\mathbb{R}^p$. Then it is clear that every L_2-integrable function has a representation of the form (*); indeed, the ordinary Fourier series representation of f is of this form.

The PP idea might be applied as follows. Assume that we already have determined projection vectors a_j and functions f_j for $j < k$. Now

determine a_k, f_k such that the norm of the residual function

$$r(x) = f(x) - \sum_{1}^{k-1} f_j\,(a_j^T x)$$

is decreased by the maximum possible amount when $f_k(a_k^T x)$ is added into the sum on the right hand side.

For fixed a_k, the solution is given by the function

$$f_k(z) = E(r(x)\,|\,a_k^T x = z),$$

where the conditional expectation is taken under the assumption that x is distributed according to the underlying probability measure F.

<u>Proof</u>. Let E' denote the conditional expectation, given $a_k^T x = z$. Then, for any function g of z,

$$E[(r-g)^2] = E\left\{E'[(r-f_k) + (f_k-g)]^2\right\}$$

$$= E\left\{E'[r-f_k)^2] + 2[E'(r-f_k)](f_k-g) + (f_k-g)^2\right\}$$

$$= E[(r-f_k)^2] + E[(f_k-g)^2] \geq E[(r-f_k)^2].$$

□

Since

$$E(r-f_k)^2 = E[E'(r-f_k)^2]$$

$$= E[E'(r^2) - 2f_k E'(r) + f_k^2]$$

$$= E[E'(r^2) - f_k^2]$$

$$= E(r^2) - E(f_k^2),$$

the decrease in the residual sum of squares is $E(f_k^2)$. The problem thus is either to maximize $E(f_k^2)$ or to minimize $E(r-f_k)^2$ through a suitable choice of a_k.

We have no reason to assume that the successive approximations

$$\tilde{f}_k(x) = \sum_{1}^{k} f_j\,(a_j^T x)$$

to f are the best possible for k summands. In general, it should be possible to improve the fit by backtracking, i.e., by omitting one of the earlier summands and determining a best possible replacement, and then iterating.

However, there are heuristic reasons to assume that such an improvement may be small and hardly worthwhile. These reasons are as follows. Assume that f is defined in the unit cube, and assume for simplicity that f integrates to 0. We interpret x modulo 1, or equivalently, we extend all functions periodically, with period 1 in each coordinate direction. Then, if we recall the Fourier expansion of f, it follows that only terms f_j in rational directions are different from 0, that is, those where a_j is a multiple of a vector $(n_1,\ldots,n_p)$ with integral components. If in the expansion (*) all a_j are different, each summand f_j then corresponds to the projection of f onto a particular rational direction, and it picks up a certain part of the Fourier expansion, namely the terms belonging to lattice points which are multiples of a_j. If the f_j are ordered according to decreasing norm, the partial sum $\sum_{1}^{k} f_j\,(a_j^T x)$ for each k clearly gives the best approximation possible for k terms.

PROBLEM 7. Investigate these matters quantitatively, in dependence of the smoothness of f.

An additive decomposition of f is by no means the only possibility. If f is a probability density, it might make more sense to decompose it multiplicatively, i.e., to approximate $\log f$ by

$$\log g(x) = \sum_{1}^{k} h_j\,(a_j^T x).$$

If $k = p$, and if the a_j are linearly independent, this means that we approximate the density f by a product measure in a suitable coordinate system. See Huber (1981b).

7. PP AND MINIMUM ENTROPY DECONVOLUTION

Assume $y = (y_1,\ldots,y_q)^T$ is an unobservable random vector with independent components, at least some of which are non-normal, and assume that the observable vector $x = (x_1,\ldots,x_p)^T$ has been generated by an unkown linear transformation applied to y:

$$x = By.$$

According to the principles of PP enunciated in Section 2, we would like to find a suitable dimension k and a least normal k-dimensional projection A. Ideally, we may hope that this will undo the unknown transformation B, so that the components of $z = Ax = (z_1,\ldots,z_k)^T$ correspond to the non-normal components of y, apart from permutations and similarity transformations.

A special case of this problem occurs in the analysis of seismic time series data. Thanks to stationarity, the y_i then are identically distributed, and A, B reduce to linear filters. Wiggins (1977, 1978) proposed to attack this problem by what he called Minimum Entropy Deconvolution (MED), but only Donoho (1980) realized that MED was in fact a special case of PP, and that, while Wiggins originally had proposed to use kurtosis as the objective function, Shannon entropy would in fact be the asymptotically optimal choice, so Wiggin's choice of terminology appears prescient. MED turns out to be the theoretically cleanest approach to the "unsmoothing" or "sharpening" problem.

PROBLEM 8. Investigate the general problem sketched at the beginning of this section. See also Kagan, Linnik and Rao, Ch. 10.

Projection pursuit, when applied to this general problem, should be able to separate the non-normal dimensions from the normal (i.e., uninteresting) ones, but it would not be able to separate out normal components lying in the space of the non-normal ones. To be more specific, let us consider the following example (with $q = 4$, $p = k = 2$, denoting the normal components among the y_i by u_i):

$$z_1 = y_1 + y_2 + u_3$$

$$z_2 = y_1 - y_2 + u_4$$

where y_1, y_2, u_3, u_4 are independent random variables, the y_i nonnormal, the u_i normal $N(0,\sigma^2)$. Then, the following transformation would be the best transformation that PP could ever hope to achieve:

$$x_1 = \frac{1}{2}(z_1 + z_2) = y_1 + \frac{1}{2}(u_3 + u_4) = y_1 + w_1,$$

$$x_2 = \frac{1}{2}(z_1 - z_2) = y_2 + \frac{1}{2}(u_3 - u_4) = y_2 + w_2,$$

where the w_i are independent normal $N(0, \frac{1}{2}\sigma^2)$. Since

$$z_1 = x_1 + x_2$$

$$z_2 = x_1 - x_2$$

this might even be considered the ultimate solution. But by grasping the idea of deconvolution (which was used in a slightly different sense

in MED, namely to deconvolute away a linear filter), we may go beyond and generalize the idea of PP: if y_i does not contain a normal part, we can reconstruct $\hat{f}(y_i)$ from $f(x_i)$ by removing a largest possible normal convolution factor.

Somewhat more generally, the idea of MED should help to analyze a p-dimensional distribution that happens to be an additive mixture (i.e., an underlying non-normal structure with superimposed noise, where the noise is the same everywhere, but not necessarily normal).

PROBLEM 9. Formalize and investigate the idea of "sharpening" implicit in the preceding paragraph.

8. THE SMOOTHING PROBLEM

So far, we operated on the underlying model, that is, we assumed that our independent variable $x \in \mathbb{R}^p$ had a probability density that was observable through its low dimensional marginal distributions. Now we assume that we only have a sample.

Then, for ordinary PP some objective functions (like kurtosis) still work without change, but some others (like Shannon entropy and Fisher information) use the marginal density and possibly its derivatives, and are no longer directly applicable; thus we need to estimate this density first. In PPR we need smoothing both to smooth away the granularity of the carrier x and to reduce the random errors in the measurements of the response $f(x)$.

Friedman and his coworkers had determined the characteristics of their smoothing procedures by heuristics plus empiricism. Complementary theoretical studies are needed, in particular on the question of bandwidth of the smoothers.

Should the bandwidth be determined by constant coverage (a fraction α of the observations) or by constant width (α times the interquartile range)? Considerations of variance stabilization would (slightly paradoxically) seem to argue in favor of estimating the squareroot of the density and using constant width. How should α depend on the sample? Investigations by Freedman and Diaconis (1981) suggest for instance that the traditional recommendations for the size of histogram bins make them too small.

The precise method of smoothing is less important--if it is computationally cheap. But care must be given to avoiding bias due to boundary effects. Should cross validation methods be used to balance bias and variability in an automatic data dependent fashion?

In general, more circumspection is needed than most of the papers in the smoothing literature do provide. A crude (say, piecewise linear) density estimate with the right bandwidth may be preferable to, say, a more refined and subjectively, better looking, spline estimate with a slightly too small bandwidth (whose fault is that the wiggles of the Brownian Bridge are not ironed away).

In first instance, the theory must provide good guidelines concerning the threshold between undersmoothing and oversmoothing--the data analyst must know on which side he is and whether he runs the risk of seeing and interpreting random ghosts, or of not seeing important structures. Of course, any such threshold is not a thin line, but has a finite extension. The traditional categories of consistency, and of asymptotic efficiency are clearly useful for finding good procedures, but they may lead astray--one should never try to polish methods down to ultimate efficiency at the cost of other, perhaps less easily quantifiable but not less important aspects, like ease of interpretation, cheap computation, or robustness.

9. PP AND COMPUTER ASSISTED TOMOGRAPHY (CAT)

This and the following section mention two topics intriguingly related to PP.

PPR, in particular a variant, PP density estimation, has the same goal as Computer Assisted Tomography (CAT): efficient reconstruction of a density from lower dimensional projections. But it adds two twists: (i) the sampling aspect, and (ii) the idea of searching for, and using only the most informative directions. The second one is of crucial importance in higher dimensions.

For literature on CAT and related topics, see Helgason (1980).

It remains to be seen whether the connections are close enough to be mutually helpful. The so-called support theorem may be the potentially most important result, cf. Helgason (1980), in particular p. 51.

10. PP AND HILBERT'S 13TH PROBLEM

In pure mathematics, the problem of the decomposition of an arbitrary function into a sum of functions of fewer variables took off from Hilbert's 13th problem. An up-to-date survey is the monograph by Vitushkin (1978).

The main result of relevance to us is a theorem of Kolmogorov, with improvements by Kahane, according to which any continuous function on the unit cube in $\mathbb{R}^p$ can be represented as a sum of $2p+1$ univariate functions.

<u>THEOREM</u>. (Vitushkin, p. 28)

Let λ_k $(k = 1,\ldots,p)$ be a collection of rationally independent constants. Then for quasi every collection $\{\phi_1,\ldots,\phi_{2p+1}\}$ of monotone continuous functions on the segment $[0,1]$ it is true that any continuous function f on $[0,1]^p$ can be represented as

$$f(x) = \sum_{\ell=1}^{2p+1} g\left(\sum_{k=1}^{p} \lambda_k\, \phi_\ell(x_k)\right). \qquad (**)$$

where g is a continuous function.

This representation tends to be a rather "wild" one, and it is not clear how it relates to the representation

$$f(x) = \sum f_k\,(a_k^T x) \qquad (*)$$

the PPR methods aim for. But in view of the Kolmogorov-Kahane theorem one wonders whether also (*) might allow an exact representation of f in a finite number of terms. Diaconis (personal communication) recently has constructed counter examples. Moreover it invites one to think of representations similar to (**) in the PP context, that is, of best possible approximations of f in terms of a sum of functions of non-linearly transformed variables, and, more generally, of otherwise non-linear projections. Compare also Tukey (1981).

REFERENCES

Blachman, N. M. (1965), The convolution inequality for entropy powers. IEEE Trans. Information Theory 11, 267-271.

Chen, Z. and Li, G. (1981), Robust principal components and dispersion matrices via projection pursuit. Research Report, Dept. of Statistics, Harvard University (in preparation).

Donoho, D. (1980), Minimum entropy deconvolution. Research Report No. Dept. of Statistics, Harvard University.

Ferguson (1961), On the rejection of outliers. Fourth Berkeley Symposium on Math. Statistics and Probability, Vol. 1, 253-288, ed. J. Neyman, University of California Press, Berkeley, CA.

Freedman, D. and Diaconis, P. (1980), On the histogram as a density estimation. Tech Rep. No. 159, Dept. of Statistics, Stanford Univ.

Friedman, J.H. (1979), A tree-structured approach to nonparametric multiple regression. In Smoothing Techniques for Curve Estimation ed. Th. Gasser and M. Rosenblatt. Lecture Notes in Mathematics No. 757, Springer-Verlag.

Friedman, J.H., Jacobson, H. and Stuetzle, W. (1980), Projection pursuit regression. Tech. Rep. No. 146, Dept. of Statistics, Stanford University.

Friedman, J.H. and Tukey, J.W. (1974), A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Computers C-23, 881-889.

Friedman, J.H. and Stuetzle, W. (1980), Projection pursuit classification (unpublished).

Harman, H.H. (1967), Modern Factor Analysis, University of Chicago Press.

Helgason S. (1980), The Radon Transform. Birkhäuser, Boston

Huber, P.J. (1981a), Robust Statistics. John Wiley & Sons, New York.

Huber, P.J. (1981b), Density estimation and projection pursuit methods (manuscript).

Kagan, A.M., Linnik, Y.V. and Rao, C.R. (1973), Characterization Problems in Mathematical Statistics. John Wiley & Sons, N.Y.

Kruskal, J.B. (1969), Toward a practical method which helps uncover the structure of a set of multivariate observations by finding a linear transformation which optimizes a new 'index of condensation', in: Statistical Computation. R.C. Milton and J.A. Nelder, Ed. New York, Academic Press.

Tukey, J.W. (1981), Control philosophy for two-handed flexible and immediate control of a graphic display. Tech. Rep. No. 197, Ser. 2, Dept. of Statistics, Princeton University.

Vitushkin, A.G. (1978), On representation of functions by means of superpositions and related topics. L'enseignement mathématique, Monographie no. 25, Geneva.

Wiggins, R.A. (1977), Minimum entropy deconvolution. Paper presented at 39th Meeting Eur. Ass. Expl. Geophysics.

Wiggins, R.A. (1978), Minimum entropy deconvolution. *Geoexploration* 16, 21-35.